4

Correlation and Regression

When it is not in our power to determine what is true, we ought to follow what is most probable.

—René Descartes

René Descartes (1596–1650)

French mathematician and philosopher.

It is important to realize that statistics and probability do not deal in the realm of certainty. If there is any realm of human knowledge in which genuine certainty exists, you may be sure that our statistical methods are not needed there. In most human endeavors, and in almost all of the natural world around us, the element of chance happenings cannot be avoided. When we cannot expect something with true certainty, we must rely on probability to be our guide. In this chapter we will study regression, correlation, and forecasting. One of the tools we use is a scatter plot. René Descartes was the first mathematician to systematically use rectangular coordinate plots. For this reason, such a coordinate axis is called a Cartesian axis.

PREVIEW QUESTIONS

- How can you use a scatter diagram to visually estimate the degree of linear correlation of two random variables? (SECTION 4.1)
- How do you compute the correlation coefficient and what does it tell you about the strength of the linear relationship between two random variables? (SECTION 4.1)
- What is the least-squares criterion? How do you find the equation of the least-squares line? (SECTION 4.2)
- What is the coefficient of determination and what does it tell you about explained variation of y in a random sample of data pairs (x, y)? (SECTION 4.2)

For on-line student resources, visit **math.college.hmco.com/students** and follow the Statistics links to the Brase/Brase, *Understanding Basic Statistics*, 4th edition web site.

4.1 Scatter Diagrams and Linear Correlation

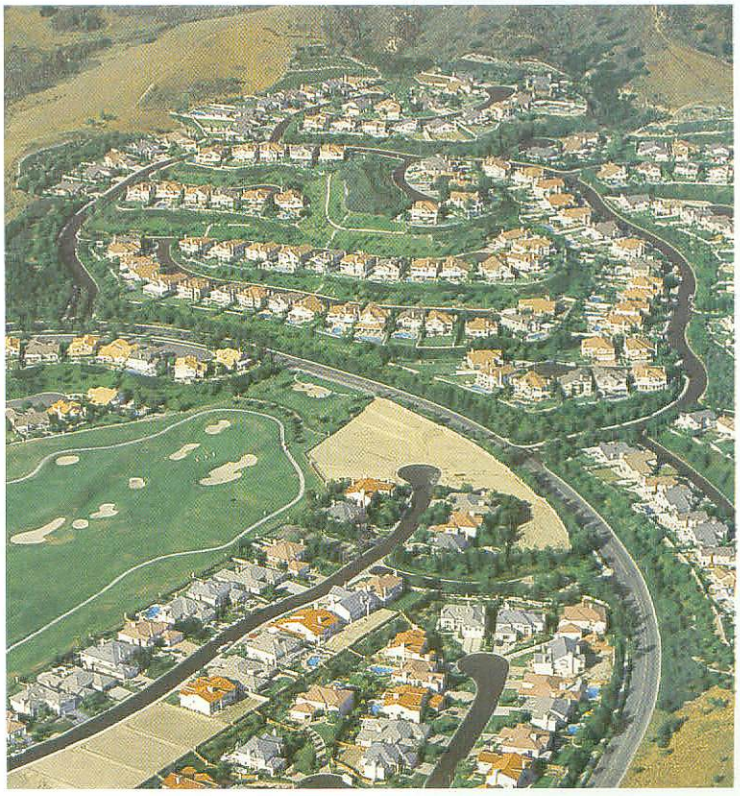
4.2 Linear Regression and the Coefficient of Determination

FOCUS PROBLEM

Changing Populations and Crime Rate

Is the crime rate higher in neighborhoods where people might not know each other very well? Is there a relationship between crime rate and population change? If so, can we make predictions based on such a relationship? Is the relationship statistically significant? Is it possible to predict crime rate from population changes?

Denver is a city that has had a lot of growth and consequently a lot of population change in recent years. Sociologists studying population changes and crime rate could find a wealth of information in Denver statistics. Let x be a random variable representing percentage change in neighborhood population in the past few years, and let y be a random variable representing crime rate (crimes per 1,000 population). A random sample of six Denver neighborhoods gave the following information (Source: *Neighborhood Facts*, The Piton Foundation). To find out more about the Piton Foundation, visit the Brase/Brase statistics site at **http://math.college.hmco.com/students** and find the link to the Piton Foundation.



x	29	2	11	17	7	6
y	173	35	132	127	69	53

Using information presented in this chapter, you will be able to analyze the relationship between the variables x and y using the following tools.

- Scatter diagram
- Sample correlation coefficient and coefficient of determination
- Least-squares line equation
- Predictions for y using the least-squares line

(See Problem 6 in the Chapter Review Problems.)

4.1 Scatter Diagrams and Linear Correlation

FOCUS POINTS

- ✓ Make a scatter diagram.
- ✓ Visually estimate the location of the "best-fitting" line for a scatter diagram.
- ✓ Use sample data to compute the sample correlation coefficient r.
- ✓ Investigate the meaning of the correlation coefficient r.

Studies of correlation and regression of two variables usually begin with a graph of *paired data values* (x, y). We call such a graph a *scatter diagram.*

A **scatter diagram** is a graph in which data pairs (x, y) are plotted as individual points on a grid with horizontal axis x and vertical axis y. We call x the **explanatory** variable and y the **response** variable.

Scatter diagram

By looking at a scatter diagram of data pairs, you can observe if there seems to be a linear relationship between the x and y values.

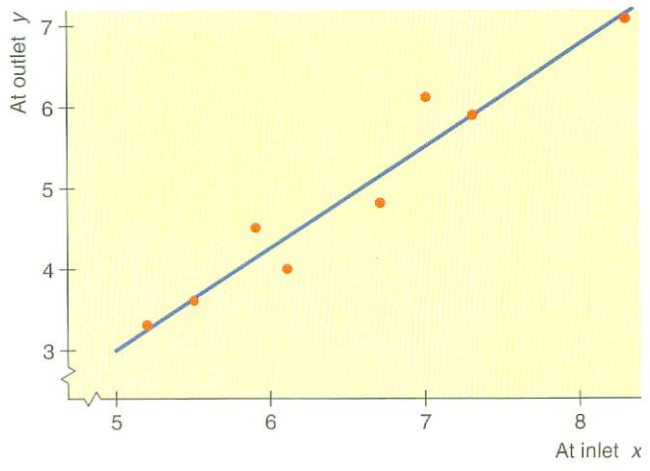
EXAMPLE 1
Scatter diagram

Phosphorous is a chemical used in many household and industrial cleaning compounds. Unfortunately, phosphorous tends to find its way into surface water, where it can kill fish, plants, and other wetland creatures. Phosphorous reduction programs are required by law and are monitored by the Environmental Protection Agency (EPA). (Reference: *EPA Case Study 832-R-93-005.*)

A random sample of eight sites in a California wetlands study gave the following information about phosphorous reduction in drainage water. In this study, x is a random variable that represents phosphorous concentration (in 100 mg/l) at the inlet of a passive biotreatment facility, and y is a random variable that represents total phosphorous concentration (in 100 mg/l) at the outlet of the passive biotreatment facility.

x	5.2	7.3	6.7	5.9	6.1	8.3	5.5	7.0
y	3.3	5.9	4.8	4.5	4.0	7.1	3.6	6.1

FIGURE 4-1



(a) Make a scatter diagram for these data.

SOLUTION: Figure 4-1 shows points corresponding to the given data pairs. These plotted points constitute the scatter diagram. To make the diagram, first scan the data and decide on an appropriate scale for each axis. Figure 4-1 shows the scatter diagram (points) along with a line segment showing the basic trend. Notice a "jump scale" on both axes.

(b) Comment on the relationship between x and y shown in Figure 4-1.

SOLUTION: By inspecting the figure, we see that smaller values of x are associated with smaller values of y and larger values of x tend to be associated with larger values of y. Roughly speaking, the general trend seems to be reasonably well represented by an upward-sloping line segment, as shown in the diagram. ◇

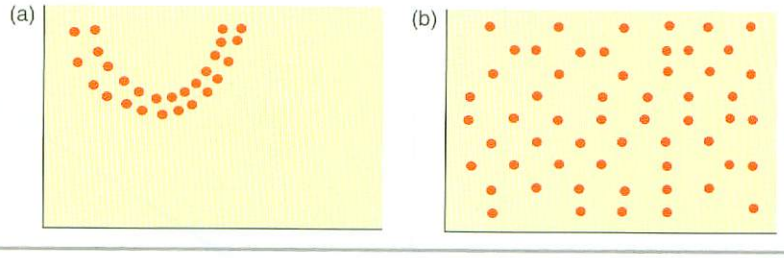
Of course, it is possible to draw many curves close to the points in Figure 4-1, but a straight line is the simplest and most widely used for elementary studies of paired data. We can draw many lines in Figure 4-1, but in some sense, the "best" line should be the one that comes closest to each of the points of the scatter diagram. To single out one line as the "best-fitting line," we must find a mathematical criterion for this line and a formula representing the line. This will be done in Section 4.2 using the *method of least squares*.

Introduction to linear correlation

Another problem precedes that of finding the "best-fitting line." That is the problem of determining how well the points of the scatter diagram are suited for

FIGURE 4-2

Scatter Diagrams with No Linear Correlation



fitting *any* line. Certainly, if the points are a very poor fit to *any* line, there is little use in trying to find the "best" line.

If the points of a scatter diagram are located so that *no* line is realistically a "good" fit, we then say that the points possess *no linear correlation*. We see some examples of scatter diagrams for which there is no linear correlation in Figure 4-2.

GUIDED EXERCISE 1

Scatter diagram

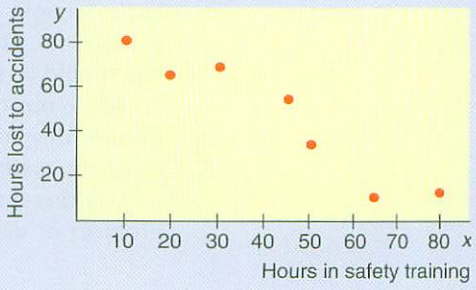
A large industrial plant has seven divisions that do the same type of work. A safety inspector visits each division of 20 workers quarterly. The number x of work-hours devoted to safety training and the number y of work-hours lost due to industry-related accidents are recorded for each separate division in Table 4-1.

TABLE 4-1 Safety Report

Division	x	y
1	10.0	80
2	19.5	65
3	30.0	68
4	45.0	55
5	50.0	35
6	65.0	10
7	80.0	12

(a) Make a scatter diagram for these pairs. Represent the x values on the horizontal axis and the y values on the vertical axis.

FIGURE 4-3 Scatter Diagram for Safety Report



Continued

GUIDED EXERCISE 1 continued

(b) As the number of hours spent on safety training increases, what happens to the number of hours lost due to industry-related accidents?
➡ In general, as the number of hours in safety training goes up, the number of hours lost due to accidents goes down.

(c) Does a line fit the data reasonably well?
➡ A line fits reasonably well.

(d) Draw a line that you think "fits best."
➡ Use a downward-sloping line that lies close to the points. Later, you will find the equation of the line that is a "best fit."

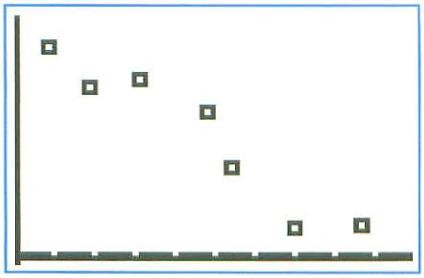
TECH NOTE The TI-84Plus and TI-83Plus calculators, Excel, and Minitab all produce scatter plots. For each technology, enter the x values in one column and the corresponding y values in another column. The displays show the data from Guided Exercise 1 regarding safety training and hours lost because of accidents. Notice that the scatter plots do not necessarily show the origin.

TI-84Plus/TI-83Plus Enter the data into two columns. Use **Stat Plot** and choose the first type. Use option **9: ZoomStat** under **Zoom.** To check the scale, look at the settings displayed under **Window.**

Excel Enter the data into two columns. Use the menu choices **Chart wizard ➤ Scatter Diagram.** Dialogue box choices permit you to label the axes and title the chart. Changing the size of the diagram box changes the scale on the axes.

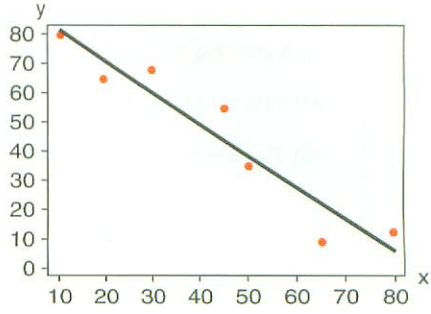
Minitab Enter the data into two columns. Use the menu selections **Stat ➤ Regression ➤ Fitted Line Plot.** The best-fit line is automatically plotted on the scatter diagram.

TI-84Plus/TI-83Plus Display



Excel Display

Minitab Display



Sample Correlation Coefficient *r*

Looking at a scatter diagram to see whether a line best describes the relationship between the values of data pairs is useful. In fact, whenever you are looking for a relationship between two variables, making a scatter diagram is a good first step.

There is a mathematical measurement that describes the strength of the linear association between two variables. This measure is the *sample correlation coefficient r.*

The full name for r is the *Pearson product-moment correlation coefficient,* named in honor of the English statistician Karl Pearson (1857–1936), who is credited with formulating r.

The **correlation coefficient** r is a numerical measurement that assesses the strength of a *linear* relationship between two variables x and y.

1. r is a unitless measurement between -1 and 1. In symbols, $-1 \leq r \leq 1$. If $r = 1$, there is perfect positive linear correlation. If $r = -1$, there is perfect negative linear correlation. If $r = 0$, there is no linear correlation. The closer r is to 1 or -1, the better a line describes the relationship between the two variables x and y.
2. Positive values of r imply that as x increases, y tends to increase. Negative values of r imply that as x increases, y tends to decrease.
3. The value of r is the same regardless of which variable is the explanatory variable and which is the response variable. In other words, the value of r is the same for the pairs (x, y) and the corresponding pairs (y, x).
4. The value of r does not change when either variable is converted to different units.

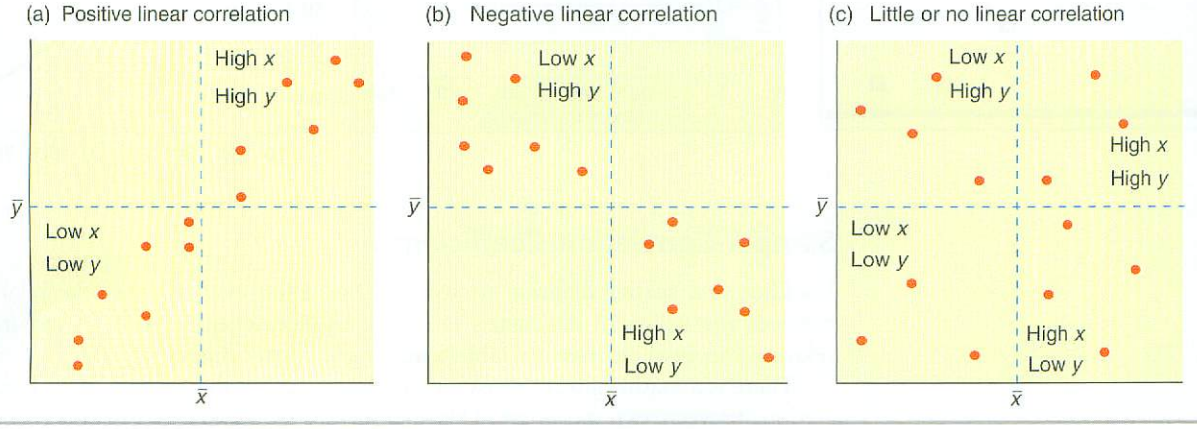
We'll develop the defining formula for r and then give a more convenient computation formula.

Development of Formula for r

If there is a *positive* linear relation between the variables x and y, then high values of x are paired with high values of y, and low values of x are paired with low values of y. [See Figure 4-4(a).] In the case of *negative* linear correlation, high values

FIGURE 4-4

Patterns for Linear Correlation



of x are paired with low values of y, and low values of x are paired with high values of y. This relation is pictured in Figure 4-4(b). If there is *little or no linear correlation* between x and y, however, then we will find both high and low x values sometimes paired with high y values and sometimes paired with low y values. This relation is shown in Figure 4-4(c).

These observations lead us to the development of the formula for the correlation coefficient r. Taking *high* to mean "above the mean," we can express the relationships pictured in Figure 4-4 by considering the products

$$(x - \bar{x})(y - \bar{y})$$

If both x and y are high, both factors will be positive, and the product will be positive as well. The sign of this product will depend on the relative values of x and y compared with their respective means.

$$(x - \bar{x})(y - \bar{y}) \begin{cases} \text{is positive if } x \text{ and } y \text{ are both "high"} \\ \text{is positive if } x \text{ and } y \text{ are both "low"} \\ \text{is negative if } x \text{ is "low," but } y \text{ is "high"} \\ \text{is negative if } x \text{ is "high," but } y \text{ is "low"} \end{cases}$$

In the case of positive linear correlation, most of the products $(x - \bar{x})(y - \bar{y})$ will be positive and so will the sum over all the data pairs

$$\Sigma(x - \bar{x})(y - \bar{y})$$

For negative linear correlation, the products will tend to be negative, so the sum also will be negative. On the other hand, in the case of little, if any, linear correlation, the sum will tend to be zero.

One trouble with the preceding sum is that it will be larger or smaller, depending on the units of x and y. Because we want r to be unitless, we standardize both x and y of a data pair by dividing each factor $(x - \bar{x})$ by the sample standard deviation s_x and each factor $(y - \bar{y})$ by s_y. Finally, we take an average of all the products. For technical reasons, we take the average by dividing by $n - 1$ instead of by n. This process leads us to the desired measurement, r.

$$r = \frac{1}{n-1} \Sigma \left(\frac{(y - \bar{y})}{s_y} \cdot \frac{(x - \bar{x})}{s_x} \right) \tag{1}$$

Computation Formula for r

The defining formula for r shows how the mean and standard deviation of each variable in the data pair enter into the formulation of r. However, the defining formula is technically difficult to work with because of all the subtractions and products. A computation formula for r uses the raw data values of x and y directly.

PROCEDURE

How to compute the sample correlation coefficient r

Obtain a random sample of n data pairs (x, y).

1. Using the data pairs, compute Σx, Σy, Σx^2, Σy^2, and Σxy.
2. With n = sample size, Σx, Σy, Σx^2, Σy^2, and Σxy, you are ready to compute the sample correlation coefficient r using the computation formula

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}} \quad (2)$$

Be careful! The notation Σx^2 means first square x and then calculate the sum, whereas $(\Sigma x)^2$ means first sum the x values, then square the result.

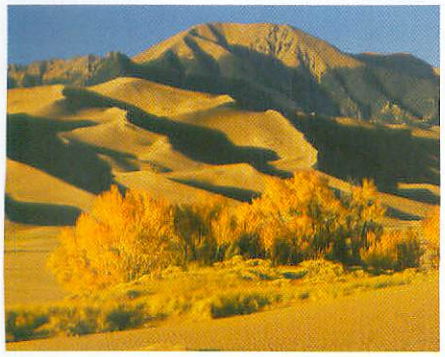
Note: Inferences for the population correlation coefficient (Section 11.4) require the data pairs to have a *bivariate normal distribution.* That is, for a fixed value of x, the y values should have a normal distribution (or at least a mound-shaped and symmetric distribution), and for a fixed value of y, the x values should have their own (approximately) normal distribution. Chapter 6 discusses normal distributions.

It can be shown mathematically that r is always a number between +1 and −1 ($-1 \le r \le +1$). Table 4-2 on the next page gives a quick summary of some basic facts about r.

For most applications you will use a calculator or computer software to compute r directly. However, to build some familiarity with the structure of the correlation coefficient, it is useful to do some calculations for yourself. Example 2 and Guided Exercise 2 show how to use the computation formula to compute r.

EXAMPLE 2

Computing r



Sand driven by wind creates large beautiful dunes at the Great Sand Dunes National Monument, Colorado. Of course, the same natural forces also create large dunes in the Great Sahara and Arabia. Is there a linear correlation between wind velocity and sand drift rate? Let x be a random variable representing wind velocity (in 10 cm/sec) and let y be a random variable representing drift rate of sand (in 100 g/cm/sec). A test site at the Great Sand Dunes National Monument gave the following information about x and y (Reference: *Hydrologic, Geologic, and Biologic Research at Great Sand Dunes National Monument,* Proceedings of the National Park Service Research Symposium).

x	70	115	105	82	93	125	88
y	3	45	21	7	16	62	12

(a) Construct a scatter diagram. Do you expect r to be positive?

SOLUTION: Figure 4-5 displays the scatter diagram. From the scatter diagram it appears that as x values increase, y values also tend to increase. Therefore, r should be positive.

TABLE 4-2 Some Facts About the Correlation Coefficient

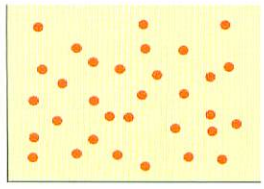
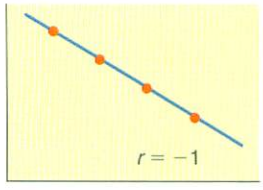
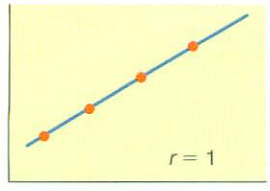
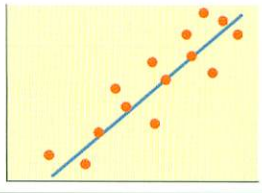
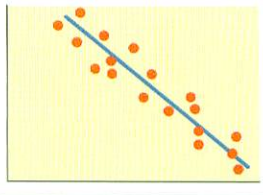
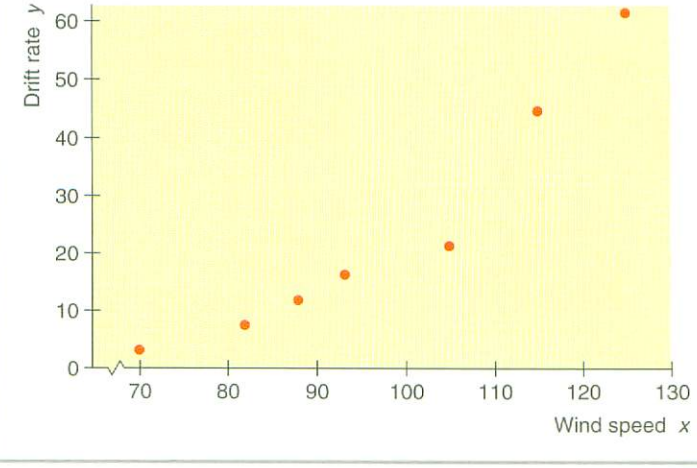
If r Is	Then	The Scatter Diagram Might Look Something Like
0	There is no linear relation among the points of the scatter diagram.	
1 or -1	There is a perfect linear relation between x and y values; all points lie on the least-squares line.	 
Between 0 and 1 ($0 < r < 1$)	The x and y values have a *positive correlation.* By this, we mean that *large* x values are associated with *large* y values, and *small* x values are associated with *small* y values.	 As we go from left to right, the least-squares line goes *up*.
Between -1 and 0 ($-1 < r < 0$)	The x and y values have a *negative correlation.* By this, we mean *large* x values are associated with *small* y values, and *small* x values are associated with *large* y values.	 As we go from left to right, the least-squares line goes *down*.

FIGURE 4-5

Wind Velocity (10 cm/sec) and Drift Rate of Sand (100 g/cm/sec)



(b) Compute r using the computation formula (formula 2).

SOLUTION: To find r, we need to compute Σx, Σx^2, Σy, Σy^2, and Σxy. It is convenient to organize the data in a table of five columns (Table 4-3) and then sum the entries in each column. Of course, many calculators give these sums directly.

TABLE 4-3 Computation Table

x	y	x^2	y^2	xy
70	3	4900	9	210
115	45	13,225	2025	5175
105	21	11,025	441	2205
82	7	6724	49	574
93	16	8649	256	1488
125	62	15,625	3844	7750
88	12	7744	144	1056
$\Sigma x = 678$	$\Sigma y = 166$	$\Sigma x^2 = 67{,}892$	$\Sigma y^2 = 6768$	$\Sigma xy = 18{,}458$

Using the computation formula for r, the sums from Table 4-3, and $n = 7$, we have

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}} \tag{2}$$

$$= \frac{7(18{,}458) - (678)(166)}{\sqrt{7(67{,}892) - (678)^2}\sqrt{7(6768) - (166)^2}} \approx \frac{16{,}658}{(124.74)(140.78)} \approx 0.949$$

Note: Using a calculator to compute r directly gives 0.949, to three places after the decimal.

(c) What does the value of r tell you?

SOLUTION: Since r is very close to 1, we have an indication of a strong positive linear correlation between wind velocity and drift rate of sand. In other words, we expect that higher wind speeds tend to mean greater drift rates. Because r is so close to 1, the association between the variables appears to be linear. ◆

It is quite a task to compute r for even seven data pairs. The use of columns as in Example 2 is extremely helpful. Your value for r should always be between -1 and 1, inclusive. Use a scatter diagram to get a rough idea of the value of r. If your computed value of r is outside the allowable range, or if it disagrees quite a bit with the scatter diagram, recheck your calculations. Be sure you distinguish between expressions such as (Σx^2) and $(\Sigma x)^2$. Negligible rounding errors may occur, depending on how you (or your calculator) round.

GUIDED EXERCISE 2

Computing r

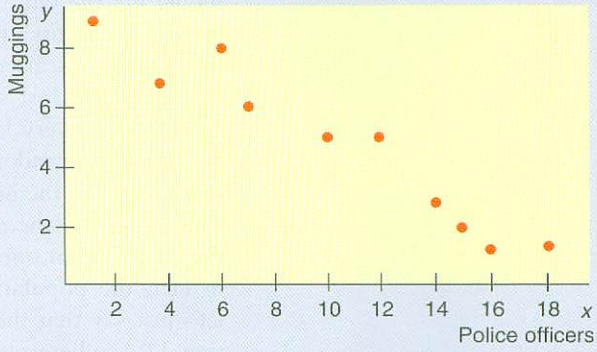
In one of the Boston city parks, there has been a problem with muggings in the summer months. A police cadet took a random sample of 10 days (out of the 90-day summer) and compiled the following data. For each day, x represents the number of police officers on duty in the park and y represents the number of reported muggings on that day.

x	10	15	16	1	4	6	18	12	14	7
y	5	2	1	9	7	8	1	5	3	6

(a) Construct a scatter diagram of x and y values.

Figure 4-6 shows the scatter diagram.

FIGURE 4-6 Scatter Diagram for Number of Police Officers versus Number of Muggings



(b) From the scatter diagram, do you think the computed value of r will be positive, negative, or zero? Explain.

r will be negative. The general trend is that large x values are associated with small y values and vice versa. From left to right, the least-squares line goes down.

(c) Verify that $\Sigma x = 103$, $\Sigma y = 47$, $\Sigma x^2 = 1347$, $\Sigma y^2 = 295$, and $\Sigma xy = 343$.

Use a calculator.

(d) Compute r. Alternatively, find the value of r directly by using a calculator or computer software.

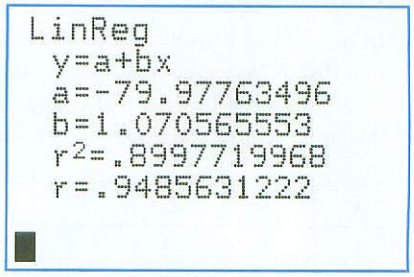
$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

$$= \frac{10(343) - (103)(47)}{\sqrt{10(1347) - (103)^2}\sqrt{10(295) - (47)^2}}$$

$$\approx \frac{-1411}{(53.49)(27.22)} \approx -0.969$$

TECH NOTE Most calculators that support two-variable statistics provide the value of the correlation coefficient r directly. Statistical software provides r, r^2, or both.

TI-84Plus/TI-83Plus First use **CATALOG**, scroll to **DiagnosticOn**, and press **Enter** twice. Then, when you use **STAT, CALC**, option **8:LinReg(a+bx)**, the value of r will be given (data from Example 2). In the next section we will discuss the line $y = a + bx$ and the meaning of r^2.



Excel Use the menu selection **Paste function** (f_x) ➤ **Statistical** ➤ **Correl.**

Minitab Use the menu selection **Stat** ➤ **Basic Statistics** ➤ **Correlation.**

Cautions about Correlation

Sample compared to population correlation

The correlation coefficient can be thought of as a measure of how well a linear model fits the data points on a scatter diagram. The closer r is to $+1$ or -1, the better a line "fits" the data. Values of r close to 0 indicate a poor fit to any line.

Usually a scatter diagram does not contain *all* possible data points that could be gathered. Most scatter diagrams represent only a *random sample* of data pairs taken from a very large population of all possible pairs. Because r is computed on the basis of a random sample of (x, y) pairs, we expect the values of r to vary from one sample to the next (much as the sample mean $\bar{x}$ varies from sample to sample). This brings up the question of the *significance* of r. Or, put another way, what are the chances that our random sample of data pairs indicates a high correlation when, in fact, the population x and y values are not so strongly correlated. Right now let's just say that the significance of r is a separate issue that will be treated in Section 11.4, where we test the *population correlation coefficient* ρ (Greek letter *rho*, pronounced "row").

r = **sample** correlation coefficient computed from a random sample of (x, y) data pairs.

ρ = **population** correlation coefficient computed from all population data pairs (x, y).

There is a less formal way to address the significance of r using a table of "critical values" or "cut-off values" based on the r distribution and the number of data pairs. Problem 15 at the end of this section discusses this method.

Causation

The correlation coefficient is a mathematical tool for measuring the strength of a linear relationship between two variables. As such, it makes no implication about cause or effect. The fact that two variables tend to increase or decrease together does not mean that a change in one is *causing* a change in the other. A strong correlation between x and y is sometimes due to other (either known or unknown) variables. Such variables are called *lurking variables*.

In ordered pairs (x, y), x is called the **explanatory** variable and y is called the **response** variable. When r indicates a linear correlation between x and y, changes in values of y tend to respond to changes in values of x according to a linear model. A **lurking variable** is a variable that is neither an explanatory nor a response variable. Yet, a lurking variable may be responsible for changes in both x and y.

EXAMPLE 3
Causation and lurking variables

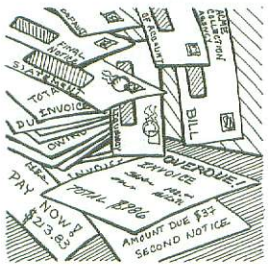
Over a period of years, the population of a certain town increased. It was observed that during this period the correlation between x, the number of people attending church, and y, the number of people in the city jail, was $r = 0.90$. Does going to church *cause* people to go to jail? Is there a *lurking variable* that might cause both variables x and y to increase?

SOLUTION: We hope church attendance does not cause people to go to jail! During this period, there was an increase in population. Therefore, it is not too surprising that both the number of people attending church and the number of people in jail increased. The high correlation between x and y is likely due to the lurking variable of population increase. ◆

Correlation between averages

The correlation between two variables consisting of averages is usually higher than the correlation between two variables representing corresponding raw data. One reason is that the use of averages reduces the variation existing between individual measurements (see Section 7.5 and the central limit theorem). A high correlation based on two variables consisting of averages does not necessarily imply a high correlation between two variables consisting of individual measurements.

VIEWPOINT



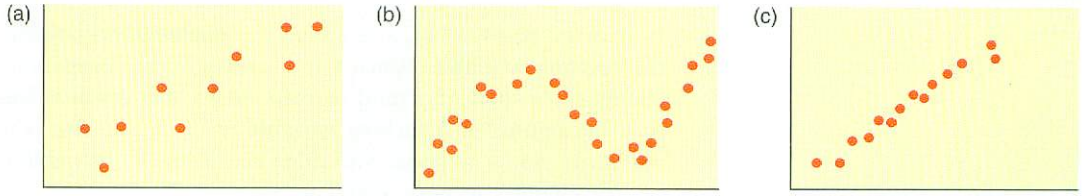
Low on Credit, High on Cost!!!

How do you measure automobile insurance risk? One way is to use a little statistics and customer credit ratings. Insurers say statistics show that drivers who have a history of bad credit are more likely to be in serious car accidents. According to a high-level executive at Allstate Insurance Company, financial instability is an extremely powerful predictor of future insurance losses. In short, there seems to be a strong correlation between bad credit ratings and auto insurance claims. Consequently, insurance companies want to charge higher premiums to customers with bad credit ratings. Consumer advocates object strongly because they say bad credit *does not cause* automobile accidents. More than 20 states prohibit or restrict the use of credit ratings to determine auto insurance premiums. Insurance companies respond by saying that your best defense is to pay your bills on time!

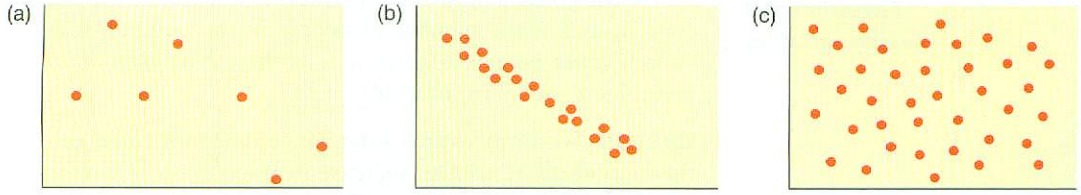
SECTION 4.1 PROBLEMS

Note: Answers may vary due to rounding.

1. *Scatter Diagrams: Linear Correlation* Look at the diagrams at the top of the next page. Does each diagram show high linear correlation, moderate or low linear correlation, or no linear correlation?



2. *Scatter Diagrams: Linear Correlation* Look at the following diagrams. Does each diagram show high linear correlation, moderate or low linear correlation, or no linear correlation?



3. *Causation: Lurking Variables* Over the past few years, there has been a strong positive correlation between the annual consumption of diet soda drinks and the number of traffic accidents.
   (a) Do you think increasing consumption of diet soda drinks causes traffic accidents? Explain.
   (b) What lurking variables might be causing the increase in one or both of the variables? Explain.

4. *Causation: Lurking Variables* Over the past decade, there has been a strong positive correlation between teacher salaries and prescription drug costs.
   (a) Do you think paying teachers more causes prescription drugs to cost more? Explain.
   (b) What lurking variables might be causing the increase in one or both of the variables? Explain.

5. *Causation: Lurking Variables* Over the past 50 years, there has been a strong negative correlation between average annual income and the record time to run 1 mile. In other words, average annual incomes have been rising while the record time to run 1 mile has been decreasing.
   (a) Do you think increasing incomes cause decreasing times to run the mile? Explain.
   (b) What lurking variables might be causing the increase or decrease in either of the variables? Explain.

6. *Causation: Lurking Variables* Over the past 30 years in the United States, there has been a strong negative correlation between the number of infant deaths at birth and the number of people over age 65.
   (a) Is the fact that people are living longer causing a decrease in infant mortalities at birth?
   (b) What lurking variables might be causing the increase or decrease in either of the variables? Explain.

7. *Veterinary Science: Shetland Ponies* How much should a healthy Shetland pony weigh? Let x be the age of the pony (in months), and let y be the average weight of the pony (in kilograms). The following information is based on data taken from *The Merck Veterinary Manual* (a reference used in most veterinary colleges).

x	3	6	12	18	24
y	60	95	140	170	185

(a) Make a scatter diagram and draw the line you think best fits the data.
(b) Would you say the correlation is low, moderate, or strong? positive or negative?
(c) Use a calculator to verify that $\Sigma x = 63$, $\Sigma x^2 = 1089$, $\Sigma y = 650$, $\Sigma y^2 = 95{,}350$, and $\Sigma xy = 9930$. Compute r. As x increases, does the value of r imply that y should tend to increase or decrease? Explain.

8. *Health Insurance: Administrative Cost* The following data are based on information from *Domestic Affairs*. Let x be the average number of employees in a group health insurance plan, and let y be the average administrative cost as a percentage of claims.

x	3	7	15	35	75
y	40	35	30	25	18

(a) Make a scatter diagram and draw the line you think best fits the data.
(b) Would you say the correlation is low, moderate, or strong? positive or negative?
(c) Use a calculator to verify that $\Sigma x = 135$, $\Sigma x^2 = 7133$, $\Sigma y = 148$, $\Sigma y^2 = 4674$, and $\Sigma xy = 3040$. Compute r. As x increases, does the value of r imply that y should tend to increase or decrease? Explain.

9. *Meteorology: Cyclones* Can a low barometer reading be used to predict maximum wind speed of an approaching tropical cyclone? Data for this problem are based on information taken from *Weatherwise* (vol. 46, no. 1), a publication of the American Meteorological Society. For a random sample of tropical cyclones, let x be the lowest pressure (in millibars) as a cyclone approaches, and let y be the maximum wind speed (in miles per hour) of the cyclone.

x	1004	975	992	935	985	932
y	40	100	65	145	80	150

(a) Make a scatter diagram and draw the line you think best fits the data.
(b) Would you say the correlation is low, moderate, or strong? positive or negative?
(c) Use a calculator to verify that $\Sigma x = 5823$, $\Sigma x^2 = 5{,}655{,}779$, $\Sigma y = 580$, $\Sigma y^2 = 65{,}750$, and $\Sigma xy = 556{,}315$. Compute r. As x increases, does the value of r imply that y should tend to increase or decrease? Explain.

10. *Geology: Earthquakes* Is the magnitude of an earthquake related to the depth below the surface at which the quake occurs? Let x be the magnitude of an earthquake (on the Richter scale), and let y be the depth (in kilometers) of the quake below the surface at the epicenter. The following is based on information taken from the National Earthquake Information Service of the U.S. Geological Survey. Additional data may be found by visiting the Brase/Brase statistics site at **http://math.college.hmco.com/students** and finding the link to earthquakes.

x	2.9	4.2	3.3	4.5	2.6	3.2	3.4
y	5.0	10.0	11.2	10.0	7.9	3.9	5.5

(a) Make a scatter diagram and draw the line you think best fits the data.
(b) Would you say the correlation is low, moderate, or strong? positive or negative?
(c) Use a calculator to verify that $\Sigma x = 24.1$, $\Sigma x^2 = 85.75$, $\Sigma y = 53.5$, $\Sigma y^2 = 458.31$, and $\Sigma xy = 190.18$. Compute r. As x increases, does the value of r imply that y should tend to increase or decrease? Explain.

11. *Archaeology: Pottery* Wind Mountain archaeological site is located in southwest New Mexico. Ancient, prehistoric pottery vessels are usually found as sherds (broken

pieces) and carefully reconstructed if enough sherds can be found. For reconstructed (or even rare unbroken) pottery vessels, let x be the body diameter (in centimeters), and let y be the height (in centimeters) of the vessel. The following data are based on information taken from *Mimbres Mogollon Archaeology*, by A. I. Woosley and A. J. McIntyre (University of New Mexico Press).

x	7.3	31.0	18.4	6.5	4.9	2.6	19.5	9.2	23.7
y	5.5	28.5	19.7	5.0	5.7	2.1	11.5	5.0	11.6

(a) Make a scatter diagram and draw the line you think best fits the data.
(b) Would you say the correlation is low, moderate, or high? positive or negative?
(c) Using a calculator, verify that $\Sigma x = 123.1$, $\Sigma x^2 = 2452.45$, $\Sigma y = 94.6$, $\Sigma y^2 = 1584.3$, and $\Sigma xy = 1897.19$. Compute r. As x increases, does the value of r imply that y should tend to increase or decrease? Explain.

12. *University Crime: FBI Report* Do larger universities tend to have more property crime? University crime statistics are affected by a variety of factors. The surrounding community, accessibility given to outside visitors, and many other factors influence crime rate. Let x be a variable that represents student enrollment (in thousands) on a university campus. Let y be a variable that represents the number of burglaries in a year on a university campus. A random sample of $n = 8$ universities in California gave the following information about enrollments and annual burglary incidents. (Reference: *Crime in the United States*, Federal Bureau of Investigation.)

x	12.5	30.0	24.5	14.3	7.5	27.7	16.2	20.1
y	26	73	39	23	15	30	15	25

(a) Make a scatter diagram and draw the line you think best fits the data.
(b) Would you say the correlation is low, moderate, or high? positive or negative?
(c) Using a calculator, verify that $\Sigma x = 152.8$, $\Sigma x^2 = 3350.98$, $\Sigma y = 246$, $\Sigma y^2 = 10{,}030$, and $\Sigma xy = 5488.4$. Compute r. As x increases, does the value of r imply that y should tend to increase or decrease? Explain.

13. *Expand Your Knowledge: Effect of Scale on Scatter Diagram* The initial visual impact of a scatter diagram depends on the scales used on the x and y axes. Consider the following data:

x	1	2	3	4	5	6
y	1	4	6	3	6	7

(a) Make a scatter diagram using the same scale on both the x and y axes (i.e., make sure the unit lengths on the two axes are equal).
(b) Make a scatter diagram using a scale on the y axis that is twice as long as that on the x axis.
(c) Make a scatter diagram using a scale on the y axis that is half as long as that on the x axis.
(d) On each of the three graphs, draw the straight line that you think best fits the data points. How do the slopes (or directions) of the three lines appear to change? (*Note:* The actual slopes will be the same; they just appear different because of the choice of scale factors.)

14. *Expand Your Knowledge: Effect on r of Exchanging x and y Values* Examine the computation formula for r, the sample correlation coefficient [formulas (1) and (2) of this section].

(a) In the formula for r, if we exchange the symbols x and y, do we get a different result or do we get the same (equivalent) result? Explain.

(b) If we have a set of x and y data values, and we exchange each corresponding x and y value to get a new data set, should the sample correlation coefficient be the same for both sets of data? Explain.

(c) Compute the sample correlation coefficient r for each of the following data sets and show that r is the same for both.

x	1	3	4
y	2	1	6

x	2	1	6
y	1	3	4

15. *Expand Your Knowledge: Using a Table to Test ρ* The correlation coefficient r is a *sample* statistic. What does it tell us about the value of the population correlation coefficient ρ (Greek letter rho)? We will build the formal structure of hypothesis tests of ρ in Section 11.4. However, there is a quick way to determine if the sample evidence based on r is strong enough to conclude that there is some population correlation between the variables. In other words, we can use the value of r to determine if $\rho \neq 0$. We do this by comparing the value $|r|$ to an entry in Table 4-4. The value of α in the table gives us the probability of concluding that $\rho \neq 0$ when in fact $\rho = 0$ and there is no population correlation. We have two choices for α: $\alpha = 0.05$ or $\alpha = 0.01$.

PROCEDURE

How to use Table 4-4 to test ρ

1. First compute r from a random sample of n data pairs (x, y).
2. Find the table entry in the row headed by n and the column headed by your choice of α. Your choice of α is the risk you are willing to take of mistakenly concluding that $\rho \neq 0$ when in fact $\rho = 0$.
3. Compare $|r|$ to the table entry.
   (a) If $|r| \geq$ table entry, then there is sufficient evidence to conclude that $\rho \neq 0$, and we say that r is **significant.** In other words, we conclude that there is some population correlation between the two variables x and y.
   (b) If $|r| <$ table entry, then the evidence is insufficient to conclude that $\rho \neq 0$, and we say that r is **not significant.** We do not have enough evidence to conclude that there is any correlation between the two variables x and y.

TABLE 4-4 Critical Values for Correlation Coefficient r

n	$\alpha = 0.05$	$\alpha = 0.01$	n	$\alpha = 0.05$	$\alpha = 0.01$	n	$\alpha = 0.05$	$\alpha = 0.01$
3	1.00	1.00	13	0.53	0.68	23	0.41	0.53
4	0.95	0.99	14	0.53	0.66	24	0.40	0.52
5	0.88	0.96	15	0.51	0.64	25	0.40	0.51
6	0.81	0.92	16	0.50	0.61	26	0.39	0.50
7	0.75	0.87	17	0.48	0.61	27	0.38	0.49
8	0.71	0.83	18	0.47	0.59	28	0.37	0.48
9	0.67	0.80	19	0.46	0.58	29	0.37	0.47
10	0.63	0.76	20	0.44	0.56	30	0.36	0.46
11	0.60	0.73	21	0.43	0.55			
12	0.58	0.71	22	0.42	0.54			

(a) Look at Problem 7 regarding the variables x = age of a Shetland pony and y = weight of that pony. Is the value of $|r|$ large enough to conclude that weight and age of Shetland ponies are correlated? Use $\alpha = 0.05$.

(b) Look at Problem 9 regarding the variables x = lowest barometric pressure as a cyclone approaches and y = maximum wind speed of the cyclone. Is the value of $|r|$ large enough to conclude that lowest barometric pressure and wind speed of a cyclone are correlated? Use $\alpha = 0.01$.

16. *Expand Your Knowledge: Sample Size and Significance of Correlation* In this problem we use Table 4-4 to explore the significance of r based on different sample sizes. See Problem 15.

(a) Is a sample correlation coefficient $r = 0.820$ significant at the $\alpha = 0.01$ level based on a sample size of $n = 7$ data pairs? What about $n = 9$ data pairs?

(b) Is a sample correlation coefficient $r = 0.40$ significant at the $\alpha = 0.05$ level based on a sample size of $n = 20$ data pairs? What about $n = 27$ data pairs?

(c) Is it true that in order to be significant, an r value must be larger than 0.90? larger than 0.70? larger than 0.50? What does sample size have to do with the significance of r? Explain.

4.2 Linear Regression and the Coefficient of Determination

FOCUS POINTS

- State the least-squares criterion.
- Use sample data to find the equation of the least-squares line. Graph the least-squares line.
- Use the least-squares line to predict a value of the response variable y for a specified value of the explanatory variable x.
- Explain the difference between interpolation and extrapolation.
- Explain why extrapolation beyond the sample data range might give results that are misleading or meaningless.
- Use r^2 to determine *explained* and *unexplained* variation of the response variable y.

In Denali National Park, Alaska, the wolf population is dependent on a large, strong caribou population. In this wild setting caribou are found in very large herds. The well-being of an entire caribou herd is not threatened by wolves. In fact, it is thought that wolves keep caribou herds strong by helping prevent overpopulation. Can the caribou population be used to predict the size of the wolf population?

Let x be a random variable that represents the fall caribou population (in hundreds) in Denali National Park. Let y be a random variable that represents the late-winter wolf population in the park. A random sample of recent years gave the following information (Reference: U.S. Department of the Interior, National Biological Service).

x	30	34	27	25	17	23	20
y	66	79	70	60	48	55	60

Looking at the scatter diagram in Figure 4-7 on the next page, we can ask some questions.

1. Do the data indicate a linear relationship between x and y?
2. Can you find an equation for the best-fitting line relating x and y? Can you use this relationship to predict the size of the wolf population when you know the size of the caribou population?
3. What fractional part of the variability in y can be associated with the variability in x? What fractional part of the variability in y is not associated with a corresponding variability in x?

FIGURE 4-7

Caribou and Wolf Populations

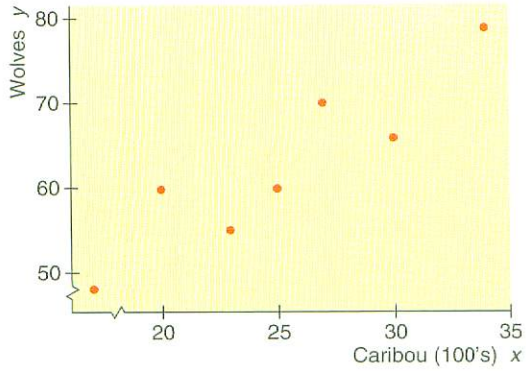
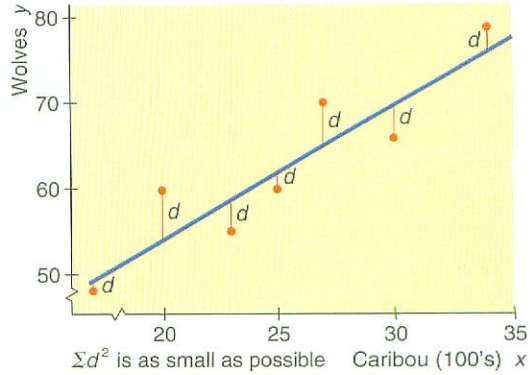


FIGURE 4-8

Least-Squares Criterion



Explanatory variable
Response variable
Least-squares criterion

The first step in answering these questions is to try to express the relationship between x and y as a mathematical equation. There are many possible equations, but the simplest and most widely used is the linear equation, or the equation of a straight line. Because we will be using this line to predict the y values from the x values, we call x the *explanatory variable* and y the *response variable.*

Our job is to find the "best" linear equation representing the points of the scatter diagram. For our criterion of best-fitting line, we use the *least-squares criterion,* which states that the line we fit to the data points must be such that *the sum of the squares of the vertical distances from the points to the line is as small as possible.* The least-squares criterion is illustrated in Figure 4-8.

Least-squares criterion

The sum of the squares of the vertical distances from the data points (x, y) to the line is made as small as possible.

In Figure 4-8, d represents the difference between the y coordinate of the data point and the corresponding y coordinate on the line. Thus, if the data point lies above the line, d is positive, but if the data point lies below the line, d is negative. As a result, the sum of the d values can be small even if the points are widely spread in the scatter diagram. However, the squares d^2 cannot be negative. By minimizing the sum of the squares, we are, in effect, not allowing positive and negative d values to "cancel out" one another in the sum. It is in this way that we can meet the least-squares criterion of minimizing the sum of the squares of the vertical distances between the points and the line over *all* points in the scatter diagram.

Least-squares line

We use the notation $\hat{y} = a + bx$ for the least-squares line. A little algebra tells us that b is the slope and a is the intercept of the line. In this context $\hat{y}$ (read

"y hat") represents the value of the response variable y estimated using the least squares line and a given value of the explanatory variable x.

Techniques of calculus can be applied to show that a and b may be computed using the following procedure.

PROCEDURE

How to find the equation for the least-squares line $\hat{y} = a + bx$

Obtain a random sample of n data pairs (x, y), where x is the *explanatory variable* and y is the *response variable.*

1. Using the data pairs, compute Σx, Σy, Σx^2, Σy^2, and Σxy. Then compute the sample means $\bar{x}$ and $\bar{y}$.
2. With n = sample size, Σx, Σy, Σx^2, Σy^2, Σxy, $\bar{x}$, and $\bar{y}$, you are ready to compute the slope b and intercept a using the computation formulas

$$\text{Slope:} \quad b = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} \tag{3}$$

$$\text{Intercept:} \quad a = \bar{y} - b\bar{x} \tag{4}$$

Be careful! The notation Σx^2 means first square x and then calculate the sum, whereas $(\Sigma x)^2$ means first sum the x values, then square the result.

3. The equation of the least-squares line computed from your sample data is

$$\hat{y} = a + bx \tag{5}$$

Note: Inferences for the population slope (Section 11.4) require the data pairs to have a *bivariate normal distribution.* That is, for a fixed value of x, the y values should have a normal distribution (or at least a mound-shaped and symmetric distribution), and for a fixed value of y, the x values should have their own (approximately) normal distribution. Chapter 6 discusses normal distributions.

◇ **COMMENT:** The computation formulas for the slope of the least-squares line, the correlation coefficient r, and the standard deviations s_x and s_y use many of the same sums. There is, in fact, a relationship between the correlation coefficient r and the slope of the least-squares line b. In instances where we know r, s_x, and s_y, we can use the following formula to compute b.

$$b = r\left(\frac{s_y}{s_x}\right) \tag{6}$$ ◇

◇ **COMMENT:** In other mathematics courses, the slope-intercept form of the equation of a line is usually given as $y = mx + b$, where m refers to the slope of the line and b to the y-coordinate of the y-intercept. In statistics, when there is only one explanatory variable, it is common practice to use the letter b to designate the slope of the least-squares line and the letter a to designate the y-coordinate of the intercept. For example, these are the symbols used on the TI-84Plus and TI-83Plus calculators as well as many other calculators. ◇

Using the formulas to find the values of a and b

For most applications, you can use a calculator or computer software to compute a and b directly. However, to build some familiarity with the structure of the computation formulas, it is useful to do some calculations for yourself. Example 4 shows how to use the computation formulas to find the values of a and b and the equation of the least-squares line $\hat{y} = a + bx$.

Note: If you are using your calculator to find the values of a and b directly, then you may omit the discussion regarding use of the formulas. Go to the margin header "Using the values of a and b to construct the equation of the least-squares line."

EXAMPLE 4
Least-squares line

Let's find the least-squares equation relating the variables x = size of caribou population (in hundreds) and y = size of wolf population in Denali National Park. Use x as the explanatory variable and y as the response variable.

(a) Use the computation formulas to find the slope of the least-squares line b and the y-intercept a.

SOLUTION: Table 4-5 gives the data values x and y, along with the values x^2, y^2, and xy. First compute the sample means.

$$\bar{x} = \frac{\Sigma x}{n} = \frac{176}{7} \approx 25.14 \quad \text{and} \quad \bar{y} = \frac{\Sigma y}{n} = \frac{438}{7} \approx 62.57$$

Next compute the slope b.

$$b = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} = \frac{7(11{,}337) - (176)(438)}{7(4628) - (176)^2} = \frac{2271}{1420} \approx 1.60$$

Use the values of b, $\bar{x}$, and $\bar{y}$ to compute the y-intercept a.

$$a = \bar{y} - b\bar{x} \approx 62.57 - 1.60(25.14) \approx 22.35$$

TABLE 4-5 Sums for Computing b, $\bar{x}$, and $\bar{y}$

x	y	x^2	y^2	xy
30	66	900	4356	1980
34	79	1156	6241	2686
27	70	729	4900	1890
25	60	625	3600	1500
17	48	289	2304	816
23	55	529	3025	1265
20	60	400	3600	1200
$\Sigma x = 176$	$\Sigma y = 438$	$\Sigma x^2 = 4628$	$\Sigma y^2 = 28{,}026$	$\Sigma xy = 11{,}337$

Using the values of a and b to construct the equation of the least-squares line

Note that calculators give the values $b \approx 1.599$ and $a \approx 22.36$. These values differ slightly from those you computed using the formulas because of rounding.

(b) Use the values of a and b (either computed or obtained from a calculator) to find the equation of the least-squares line.

SOLUTION:

$$\hat{y} = a + bx$$

$$\hat{y} \approx 22.35 + 1.60x \quad \text{since} \quad a \approx 22.35 \quad \text{and} \quad b \approx 1.60$$

Graphing the least squares line

(c) Graph the equation of the least-squares line on a scatter diagram.

SOLUTION: To graph the least-squares line, we have several options available. The slope-intercept method of algebra is probably the quickest, but may not always be convenient if the intercept is not within the range of the sample data values. It is just as easy to select two x values in the range of the x data values and then use the least-squares line to compute two corresponding $\hat{y}$ values.

In fact, we already have the coordinates of one point on the least squares line. By the formula for the intercept [Equation (4)], the point $(\bar{x}, \bar{y})$ is always on the least-squares line. For our example, $(\bar{x}, \bar{y}) = (25.14, 62.57)$.

The point $(\bar{x}, \bar{y})$ is always on the least-squares line.

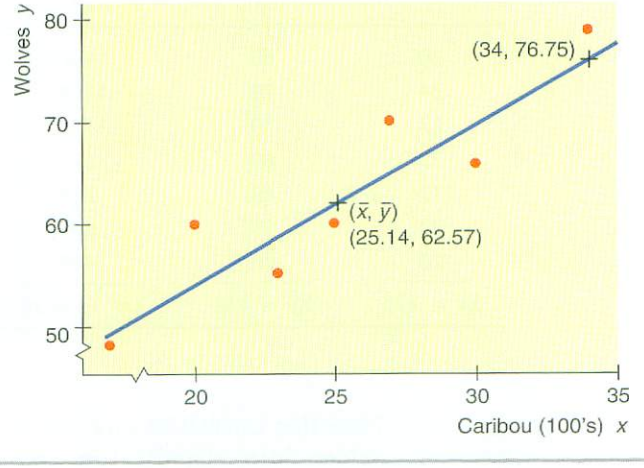
Another x value within the data range is $x = 34$. Using the least-squares line to compute the corresponding $\hat{y}$ value gives

$$\hat{y} \approx 22.35 + 1.60(34) \approx 76.75$$

We place the two points (25.14, 62.57) and (34, 76.75) on the scatter diagram (using a different symbol than that used for the sample data points) and connect the points with a line segment (Figure 4-9). ◆

FIGURE 4-9

Caribou and Wolf Populations



Meaning of slope

In the equation $\hat{y} = a + bx$, the slope b tells us how many units $\hat{y}$ changes for each unit change in x. In Example 4 regarding size of wolf and caribou populations,

$$\hat{y} \approx 22.35 + 1.60x$$

The slope 1.60 tells us that if the number of caribou (in hundreds) changes by 1 (hundred), then we expect the sustainable wolf population to change by 1.60. In other words, our model says that an increase of 100 caribou will increase the predicted wolf population by 1.60. If the caribou population decreases by 400, we predict the sustainable wolf population will decrease by 6.4.

The slope of the least-squares line tells how many units the response variable is expected to change for each unit change in the explanatory variable. The number of units change in the response variable for each unit change in the explanatory variable is called the **marginal change** of the response variable.

Predicting y for a specified x

Making predictions is one of the main applications of linear regression. In other words, you use the equation of the least-squares line to predict the $\hat{y}$ value for a specified x value. Of course, the accuracy of the prediction depends on how well the least-squares line fits the original raw data points. It is a good idea to check that the correlation coefficient indicates a strong linear correlation.

Interpolation, extrapolation

Another issue that affects the validity of predictions is whether you are *interpolating* or *extrapolating*.

Predicting $\hat{y}$ values for x values that are **between** observed x values in the data set is called **interpolation.**

Predicting $\hat{y}$ values for x values that are **beyond** observed x values in the data set is called **extrapolation.**

The least-squares line is developed from sample data pairs (x, y). The least-squares line may not reflect the relationship between x and y for values of x outside the data range. For example, there is a fairly high correlation between height and age for boys ages 1 year to 10 years. In general, the older the boy, the taller the boy. A least-squares line based on such data would give good predictions of height for ages between 1 and 10. However, it would be fairly meaningless to use the same linear regression line to predict the height of a 20-year-old or 50-year-old man.

Another consideration when working with predictions is the fact that the least-squares line is based on sample data. Each different sample will produce a slightly different equation for the least-squares line.

One more important fact about predictions. The least-squares line is developed with x as the explanatory variable and y as the response variable. This model can

be used only to predict y values from specified x values. If you wish to begin with y values and predict corresponding x values, you must start all over and compute a new equation. Such an equation would be developed using a model with x as the response variable and y as the explanatory variable. See Problem 11 at the end of this section. Note that the equation for predicting x values *cannot* be derived from the least-squares line predicting y simply by solving the equation for x.

The least-squares line developed with x as the explanatory variable and y as the response variable can be used only to predict y values from specified x values.

The next example shows how to use the least-squares line for predictions.

EXAMPLE 5
Predictions

We continue with Example 4 regarding size of the wolf population as it relates to size of the caribou population. Suppose you want to predict the size of the wolf population when the size of the caribou population is 21 (hundred).

(a) In the least-squares model developed in Example 4, which is the explanatory variable and which is the response variable? Can you use the equation to predict the size of the wolf population for a specified size of caribou population?

SOLUTION: The least-squares line $\hat{y} \approx 22.35 + 1.60x$ was developed using x = size of caribou population in hundreds as the explanatory variable and y = size of wolf population as the response variable. We can use the equation to predict the y value for a specified x value.

(b) The sample data pairs have x values ranging from 17 (hundred) to 34 (hundred) for the size of the caribou population. To predict the size of the wolf population when the size of the caribou population is 21 (hundred), will you be interpolating or extrapolating?

SOLUTION: Interpolating, since 21 (hundred) falls within the range of sample x values.

(c) Predict the size of the wolf population when the caribou population is 21 (hundred).

SOLUTION: Using the least-squares line from Example 4 and the value 21 in place of x gives

$$\hat{y} \approx 22.35 + 1.60x \approx 22.35 + 1.60(21) \approx 55.95$$

Rounding up to a whole number gives a prediction of 56 for the size of the wolf population. ◆

GUIDED EXERCISE 3

Least-squares line

The Quick Sell car dealership has been using 1-minute spot ads on a local TV station. The ads always occur during the evening hours and advertise the different models and price ranges of cars on the lot that week. During a 10-week period, the Quick Sell dealer kept a weekly record of the number x of TV ads versus the number y of cars sold. The results are given in Table 4-6.

The manager decided that Quick Sell can afford only 12 ads per week. At that level of advertisement, how many cars can Quick Sell expect to sell each week? We'll answer this question in several steps.

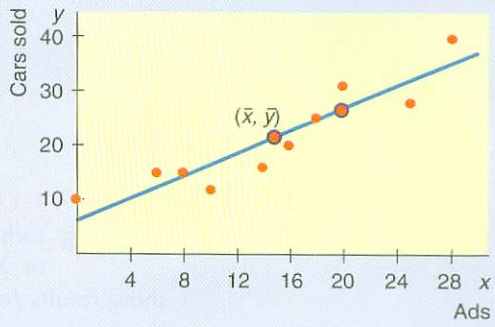
TABLE 4-6

x	y
6	15
20	31
0	10
14	16
25	28
16	20
28	40
18	25
10	12
8	15

(a) Draw a scatter diagram for the data.

The scatter diagram is shown in Figure 4-10. The plain red dots in Figure 4-10 are the points of the scatter diagram. Notice that the least-squares line is also shown with two extra points used to position the line.

FIGURE 4-10 Scatter Diagram and Least-Squares Line for Table 4-6



(b) Verify that $\Sigma x = 145$, $\Sigma y = 212$, $\Sigma x^2 = 2785$, and $\Sigma xy = 3764$.

Use a calculator.

(c) Compute the sample means $\bar{x}$ and $\bar{y}$.

$$\bar{x} = \frac{\Sigma x}{n} = \frac{145}{10} = 14.5; \quad \bar{y} = \frac{\Sigma y}{n} = \frac{212}{10} = 21.2$$

Continued

GUIDED EXERCISE 3 continued

(d) Compute a and b for the equation $\hat{y} = a + bx$ of the least-squares line.

$$b = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$$

$$= \frac{10(3764) - (145)(212)}{10(2785) - (145)^2} = \frac{6900}{6825} \approx 1.01$$

$$a = \bar{y} - b\bar{x}$$

$$\approx 21.2 - 1.01(14.5) \approx 6.56$$

(e) What is the equation of the least-squares line $\hat{y} = a + bx$?

Using the values of a and b computed in part (d) or values of a and b obtained directly from a calculator,

$$\hat{y} \approx 6.56 + 1.01x$$

(f) Plot the least-squares line on your scatter diagram.

The least-squares line goes through the point $(\bar{x}, \bar{y}) = (14.5, 21.2)$. To get another point on the line, select a value for x and compute the corresponding y value using the equation $y = 6.56 + 1.01x$. For $x = 20$, we get $y = 6.56 + 1.01(20) = 26.8$, so the point (20, 26.8) is also on the line. The least-squares line is shown in Figure 4-10.

(g) Read the y value for $x = 12$ from your graph. Then use the equation of the least-squares line to calculate y when $x = 12$. How many cars can the manager expect to sell if 12 ads per week are aired on TV?

The graph gives $y \approx 19$. From the equation we get

$$y = 6.56 + 1.01x$$

$$= 6.56 + 1.01(12) \quad \text{using 12 in place of } x$$

$$= 18.68$$

To the nearest whole number, the manager can expect to sell 19 cars when 12 ads are aired on TV each week.

TECH NOTE When we have more data pairs, it is convenient to use a technology tool such as the TI-84Plus and TI-83Plus calculators, Excel, or Minitab to find the equation of the least-squares line. The displays show results for the data of Guided Exercise 3 regarding car sales and ads.

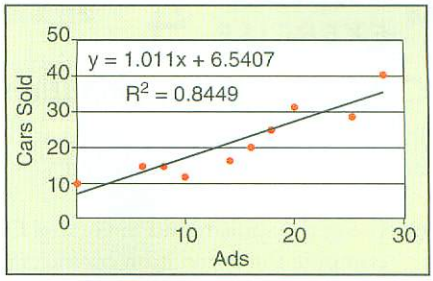
TI-84Plus/TI-83Plus Press **STAT**, choose **Calculate**, and use option **8:LinReg(a+bx)**. For a graph showing the scatter plot and the least-squares line, press the **STAT PLOT** key, turn on a plot, and highlight the first type. Then press the **Y=** key. To enter the equation of the least-squares line, press **VARS**, select **5:Statistics**, highlight **EQ**, and then select **1:RegEQ**. Press **ENTER**. Finally, press **ZOOM** and choose **9:ZoomStat**.

Excel There are several ways to find the equation of the least-squares line in Excel. One way is to make a scatter plot using the menu choices **Chart wizard ➤ Scatter Diagram**. When the diagram is complete, **right click** on one of the points on the diagram, select **trendline**, and under options check to display the equation of the line.

TI-84Plus/TI-83Plus Display

```
LinReg
 y=a+bx
 a=6.5407
 b=1.0110
 r²=.8449
 r=.9192
```

Excel Display



Minitab There are a number of ways to generate the least-squares line. One way is to use the menu selection **Stat ➤ Regression ➤ Fitted Line Plot.** The least-squares equation is shown with the diagram.

Coefficient of Determination

There is another way to answer the question, How good is the least-squares line as an instrument of regression? The *coefficient of determination* r^2 is the square of the sample correlation coefficient r.

Coefficient of determination r^2

1. Compute the sample correlation coefficient r using the procedure of Section 4.1. Then simply compute r^2, the sample coefficient of determination.
2. The value r^2 is the ratio of explained variation over total variation. That is, r^2 is the fractional amount of total variation in y that can be explained by using the linear model $\hat{y} = a + bx$.
3. Furthermore, $1 - r^2$ is the fractional amount of total variation in y that is due to random chance or to the possibility of lurking variables that influence y.

In other words, the coefficient of determination r^2 is a measure of the proportion of variation in y that is explained by the regression line, using x as the explanatory variable. If $r = 0.90$, then $r^2 = 0.81$ is the coefficient of determination. We can say that about 81% of the (variation) behavior of the y variable can be explained by the corresponding (variation) behavior of the x variable if we use the equation of the least-squares line. The remaining 19% of the (variation) behavior of the y variable is due to random chance or to the possibility of lurking variables that influence y.

GUIDED EXERCISE 4

Coefficient of determination r^2

In Guided Exercise 3 we looked at the relationship between x = number of 1-minute spot ads on TV advertising different models of cars and y = number of cars sold each week by the sponsoring car dealership.

(a) Using the sums found in Guided Exercise 3, compute the correlation coefficient r. $n = 10$, $\Sigma x = 145$, $\Sigma y = 212$, $\Sigma x^2 = 2785$, and $\Sigma xy = 3764$. You also need $\Sigma y^2 = 5320$.

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

➡ $$r = \frac{10(3764) - (145)(212)}{\sqrt{10(2785) - (145)^2}\sqrt{10(5320) - (212)^2}}$$
$$\approx \frac{6900}{(82.61)(90.86)}$$
$$\approx 0.919$$

(b) Compute the coefficient of determination r^2. ➡ $r^2 \approx 0.845$

(c) What percentage of the variation in the number of car sales can be explained by the ads and the least-squares line? ➡ 84.5%

(d) What percentage of the variation in the number of car sales is not explained by the ads and the least-squares line? ➡ 100% − 84.5%, or 15.5%

VIEWPOINT

It's Freezing!

Can you use average temperatures in January to predict how bad the rest of the winter will be? Can you predict the number of days with freezing temperatures for the entire calendar year using conditions in January? How good would such a forecast be for predicting growing season or number of frost-free days? Methods of this section can help you answer such questions. For more information, visit the Brase/Brase statistics site at **http://math.college.hmco.com/students** and find the link to temperatures.

SECTION 4.2 PROBLEMS

For Problems 1–8, please do the following.

(a) Draw a scatter diagram displaying the data.

(b) Verify the given sums Σx, Σy, Σx^2, Σy^2, and Σxy and the value of the sample correlation coefficient r.

(c) Find $\bar{x}$, $\bar{y}$, a, and b. Then find the equation of the least squares line $\hat{y} = a + bx$.

(d) Graph the least-squares line on your scatter diagram. Be sure to use the point $(\bar{x}, \bar{y})$ as one of the points on the line.

(e) Find the value of the coefficient of determination r^2. What percentage of the variation in y can be *explained* by the corresponding variation in x and the least-squares line? What percentage is *unexplained*?

Answers may vary slightly due to rounding.

1. *Economics: Entry-Level Jobs* An economist is studying the job market in Denver area neighborhoods. Let x represent the total number of jobs in a given neighborhood, and let y represent the number of entry-level jobs in the same neighborhood. A sample of six Denver neighborhoods gave the following information (units in 100s of jobs).

x	16	33	50	28	50	25
y	2	3	6	5	9	3

Source: Neighborhood Facts, The Piton Foundation. To find out more, visit the Brase/Brase statistics site at **http://math.college.hmco.com/students** and find the link to the Piton Foundation.

Complete parts (a) through (e), given $\Sigma x = 202$, $\Sigma y = 28$, $\Sigma x^2 = 7754$, $\Sigma y^2 = 164$, $\Sigma xy = 1096$, and $r \approx 0.860$.

(f) For a neighborhood with $x = 40$ jobs, how many are predicted to be entry-level jobs?

2. *Ranching: Cattle* You are the foreman of the Bar-S cattle ranch in Colorado. A neighboring ranch has calves for sale, and you are going to buy some calves to add to the Bar-S herd. How much should a healthy calf weigh? Let x be the age of the calf (in weeks), and let y be the weight of the calf (in kilograms). The following information is based on data taken from *The Merck Veterinary Manual* (a reference used by many ranchers).

x	1	3	10	16	26	36
y	42	50	75	100	150	200

Complete parts (a) through (e), given $\Sigma x = 92$, $\Sigma y = 617$, $\Sigma x^2 = 2338$, $\Sigma y^2 = 82{,}389$, $\Sigma xy = 13{,}642$, and $r \approx 0.998$.

(f) The calves you want to buy are 12 weeks old. What does the least-squares line predict for a healthy weight?

3. *Weight of Car: Miles per Gallon* Do heavier cars really use more gasoline? Suppose that a car is chosen at random. Let x be the weight of the car (in hundreds of pounds), and let y be the miles per gallon (mpg). The following information is based on data taken from *Consumer Reports* (vol. 62, no. 4).

x	27	44	32	47	23	40	34	52
y	30	19	24	13	29	17	21	14

Complete parts (a) through (e), given $\Sigma x = 299$, $\Sigma y = 167$, $\Sigma x^2 = 11{,}887$, $\Sigma y^2 = 3773$, $\Sigma xy = 5814$, and $r \approx -0.946$.

(f) Suppose that a car weighs $x = 38$ (hundred pounds). What does the least-squares line forecast for y = miles per gallon?

4. *Basketball: Fouls* Data for this problem are based on information from *STATS Basketball Scoreboard.* It is thought that basketball teams that make too many fouls in

a game tend to lose the game even if they otherwise play well. Let x be the number of fouls more than (i.e., over and above) the opposing team. Let y be the percentage of times the team with the larger number of fouls wins the game.

x	0	2	5	6
y	50	45	33	26

Complete parts (a) through (e), given $\Sigma x = 13$, $\Sigma y = 154$, $\Sigma x^2 = 65$, $\Sigma y^2 = 6290$, $\Sigma xy = 411$, and $r \approx -0.988$.

(f) If a team had $x = 4$ fouls over and above the opposing team, what does the least-squares equation forecast for y?

5. *Auto Accidents: Age* Data for this problem are based on information taken from *The Wall Street Journal*. Let x be the age in years of a licensed automobile driver. Let y be the percentage of all fatal accidents (for a given age) due to speeding. For example, the first data pair indicates that 36% of all fatal accidents involving 17-year-olds are due to speeding.

x	17	27	37	47	57	67	77
y	36	25	20	12	10	7	5

Complete parts (a) through (e), given $\Sigma x = 329$, $\Sigma y = 115$, $\Sigma x^2 = 18{,}263$, $\Sigma y^2 = 2639$, $\Sigma xy = 4015$, and $r \approx -0.959$.

(f) Predict the percentage of all fatal accidents due to speeding for 25-year-olds.

6. *Auto Accidents: Age* Let x be the age of a licensed driver in years. Let y be the percentage of all fatal accidents (for a given age) due to failure to yield the right of way. For example, the first data pair says that 5% of all fatal accidents involving 37-year-olds are due to failure to yield the right of way. *The Wall Street Journal* article referenced in Problem 5 reported the following data:

x	37	47	57	67	77	87
y	5	8	10	16	30	43

Complete parts (a) through (e), given $\Sigma x = 372$, $\Sigma y = 112$, $\Sigma x^2 = 24{,}814$, $\Sigma y^2 = 3194$, $\Sigma xy = 8254$, and $r \approx 0.943$.

(f) Predict the percentage of all fatal accidents due to failing to yield the right of way for 70-year-olds.

7. *Archaeology: Artifacts* Data for this problem are based on information taken from *Prehistoric New Mexico: Background for Survey* (by D. E. Stuart and R. P. Gauthier, University of New Mexico Press). It is thought that prehistoric Indians did not take their best tools, pottery, and household items when they visited higher elevations for their summer camps. It is hypothesized that archaeological sites tend to lose their cultural identity and specific cultural affiliation as the elevation of the site increases. Let x be the elevation (in thousands of feet) for an archaeological site in the southwestern United States. Let y be the percentage of unidentified artifacts (no

specific cultural affiliation) at a given elevation. The following data were obtained for a collection of archaeological sites in New Mexico:

x	5.25	5.75	6.25	6.75	7.25
y	19	13	33	37	62

Complete parts (a) through (e), given $\Sigma x = 31.25$, $\Sigma y = 164$, $\Sigma x^2 \approx 197.813$, $\Sigma y^2 = 6832$, $\Sigma xy = 1080$, and $r \approx 0.913$.

(f) At an archaeological site with elevation 6.5 (thousand feet), what does the least-squares equation forecast for y = percentage of culturally unidentified artifacts?

8. *Cricket Chirps: Temperature* Anyone who has been outdoors on a summer evening has probably heard crickets. Did you know that it is possible to use the cricket as a thermometer? Crickets tend to chirp more frequently as temperatures increase. This phenomenon was studied in detail by George W. Pierce, a physics professor at Harvard. In the following data, x is a random variable representing chirps per second and y is a random variable representing temperature (°F). These data are on the statSpace CD-ROM.

x	20.0	16.0	19.8	18.4	17.1	15.5	14.7	17.1
y	88.6	71.6	93.3	84.3	80.6	75.2	69.7	82.0

x	15.4	16.2	15.0	17.2	16.0	17.0	14.4
y	69.4	83.3	79.6	82.6	80.6	83.5	76.3

Source: Reprinted by permission of the publisher from *The Songs of Insects* by George W. Pierce, Cambridge, Mass.: Harvard University Press, Copyright © 1948 by the President and Fellows of Harvard College.

Complete parts (a) through (e), given $\Sigma x = 249.8$, $\Sigma y = 1200.6$, $\Sigma x^2 = 4200.56$, $\Sigma y^2 = 96{,}725.86$, $\Sigma xy = 20{,}127.47$, and $r \approx 0.835$.

(f) What is the predicted temperature when $x = 19$ chirps per second?

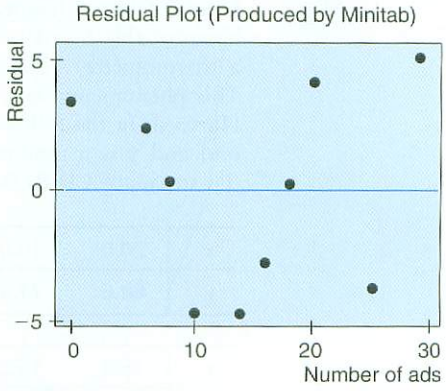
9. *Expand Your Knowledge: Residual Plot* The least-squares line usually does not go through all the sample data points (x, y). In fact, for a specified x value from a data pair (x, y), there is usually a difference between the predicted value $\hat{y}$ and the y value paired with x. This difference is called the *residual.*

The **residual** is the difference between the y value in a specified data pair (x, y) and the value $\hat{y} = a + bx$ predicted by the least-squares line for the same x.

$y - \hat{y}$ is the **residual.**

One way to assess how well a least-squares line serves as a model for the data is a **residual plot.** To make a residual plot, we put the x values in order on the horizontal axis and plot the corresponding residuals $y - \hat{y}$ in the vertical direction. Because for a least-squares model the mean of the residuals is always zero, we dash in a horizontal line at zero. The figure on the next page shows a residual plot for the data of Guided Exercise 3, in which the relationship between the number of ads run per week and the number of cars sold that week was explored. To make the residual plot, first compute all the residuals. Remember that x and y are the given data values, and $\hat{y}$ is computed from the least-squares line $\hat{y} \approx 6.56 + 1.01x$.

x	y	$\hat{y}$	Residual $y - \hat{y}$	x	y	$\hat{y}$	Residual $y - \hat{y}$
6	15	12.6	2.4	16	20	22.7	−2.7
20	31	26.8	4.2	28	40	34.8	5.2
0	10	6.6	3.4	18	25	24.7	0.3
14	16	20.7	−4.7	10	12	16.7	−4.7
25	28	31.8	−3.8	8	15	14.6	0.4



(a) If the least-squares line provides a reasonable model for the data, the pattern of points in the plot will seem random and unstructured around the horizontal line at 0. Is this the case for the residual plot?

(b) If a point on the residual plot seems far outside the pattern of other points, it might reflect an unusual data point (x, y), called an *outlier*. Such points may have quite an influence on the least-squares model. Do there appear to be any outliers in the data for the residual plot?

10. *Residual Plot: Miles per Gallon* Consider the data of Problem 3.
(a) Make a residual plot for the least-squares model.
(b) Use the residual plot to comment about the appropriateness of the least-squares model for these data. See Problem 9.

11. *Least-Squares Equation: Exchange x and y*
(a) Suppose that you are given the following x, y data pairs:

x	1	3	4
y	2	1	6

Show that the least-squares equation for these data is $y = 1.071x + 0.143$ (rounded to three digits after the decimal).

(b) Now suppose that you are given these x, y data pairs:

x	2	1	6
y	1	3	4

Show that the least-squares equation for these data is $y = 0.357x + 1.595$ (rounded to three digits after the decimal).

(c) In the data for parts (a) and (b), did we simply exchange the x and y values of each data pair?

(d) Solve $y = 0.143 + 1.071x$ for x. Do you get the least-squares equation of part (b) with the symbols x and y exchanged?

(e) In general, suppose that we have the least-squares equation $y = a + bx$ for a set of data pairs x and y. If we solve this equation for x, will we *necessarily* get the least-squares equation for the set of data pairs y, x (with x and y exchanged)? Explain using parts (a) through (d).

SUMMARY

Scatter diagrams of data pairs (x, y) are useful in helping us determine visually if there is any relation between x and y values and, if so, how strong the relation might be. We call x the explanatory variable and y the response variable.

The Pearson product-moment correlation coefficient r gives a numerical measurement assessing the strength of a *linear* relationship between x and y based on a random sample of data pairs (x, y). The value of r ranges from -1 to 1, with 1 indicating perfect positive linear correlation, -1 indicating perfect negative linear correlation, and 0 indicating no linear correlation. The closer the sample statistic r is to 1 or -1, the stronger the linear correlation.

If the scatter diagram and correlation coefficient r indicate a linear relationship, then we use the least-squares criterion to develop the equation of the least-squares line between the explanatory variable x and the response variable y

$$\hat{y} = a + bx$$

where $\hat{y}$ is the value of y predicted by the least-squares line, a is the y-intercept, and b is the slope.

The coefficient of determination r^2 is a value that measures the proportion of variation in y explained by the least-squares line.

IMPORTANT WORDS & SYMBOLS

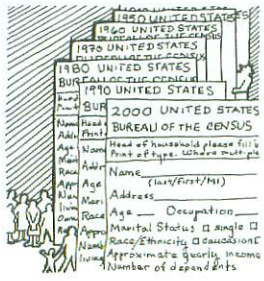
Section 4.1

Paired data values
Explanatory variable
Response variable
Scatter diagram
Sample correlation coefficient r
Perfect linear correlation
No linear correlation
Lurking variable
Meaning of slope
Interpolation
Extrapolation
Explained variation
Unexplained variation
Coefficient of determination r^2
Residual
Residual plot

Section 4.2

Least-squares criterion
Least-squares line $\hat{y} = a + bx$

VIEWPOINT



Living Arrangements

Male, female, married, single, living alone, living with friends or relatives—all these categories are of interest to the U.S. Census Bureau. In addition to these categories, there are others, such as age, income, and health needs. How strongly correlated are these variables? Can we use one or more of these variables to predict the others? How good is such a prediction? Methods of this chapter can help you answer such questions. For more information regarding such data, visit the Brase/Brase statistics site at **http://math.college.hmco.com/students** and find the link to Census Bureau.

CHAPTER REVIEW PROBLEMS

For Problems 1–6, please do the following.

(a) Draw a scatter diagram.

(b) Find the equation of the least-squares line, and plot the line on the scatter diagram of part (a).

(c) Find the correlation coefficient r. Find the coefficient of determination r^2. What percentage of variation in y is explained by the variation in x and the least-squares model?

1. *Desert Ecology: Wildlife* Bighorn sheep are beautiful wild animals found throughout the western United States. Data for this problem are based on information taken from *The Desert Bighorn*, edited by Monson and Sumner (University of Arizona Press). Let x be the age of a bighorn sheep (in years), and let y be the mortality rate (percent that die) for this age group. For example, $x = 1$, $y = 14$ means that 14% of the bighorn sheep between 1 and 2 years old died. A random sample of Arizona bighorn sheep gave the following information:

x	1	2	3	4	5
y	14	18.9	14.4	19.6	20.0

$\Sigma x = 15$; $\Sigma y = 86.9$; $\Sigma x^2 = 55$; $\Sigma y^2 = 1544.73$; $\Sigma xy = 273.4$

2. *Sociology: Job Changes* A sociologist is interested in the relation between x = number of job changes and y = annual salary (in thousands of dollars) for people living in the Nashville area. A random sample of 10 people employed in Nashville provided the following information:

x (Number of job changes)	4	7	5	6	1	5	9	10	10	3
y (Salary in $1000)	33	37	34	32	32	38	43	37	40	33

$\Sigma x = 60$; $\Sigma y = 359$; $\Sigma x^2 = 442$; $\Sigma y^2 = 13{,}013$; $\Sigma xy = 2231$

(d) If someone had $x = 2$ job changes, what does the least-squares line predict for y, the annual salary?

3. *Medical: Fat Babies* Modern medical practice tells us not to encourage babies to become too fat. Is there a positive correlation between the weight x of a 1-year-old baby and the weight y of the mature adult (30 years old)? A random sample of medical files produced the following information for 14 females:

x (lb)	21	25	23	24	20	15	25	21	17	24	26	22	18	19
y (lb)	125	125	120	125	130	120	145	130	130	130	130	140	110	115

$\Sigma x = 300$; $\Sigma y = 1775$; $\Sigma x^2 = 6572$; $\Sigma y^2 = 226{,}125$; $\Sigma xy = 38{,}220$

(d) If a female baby weighs 20 lb at 1 year, what do you predict she will weigh at 30 years of age?

4. *Sales: Insurance* Dorothy Kelly sells life insurance for the Prudence Insurance Company. She sells insurance by making visits to her clients' homes. Dorothy believes that the number of sales should depend, to some degree, on the number of visits made. For the past several years, she has kept careful records of the number of visits (x) she made each week and the number of people (y) who bought insurance that week. For a random sample of 15 such weeks, the x and y values follow:

x	11	19	16	13	28	5	20	14	22	7	15	29	8	25	16
y	3	11	8	5	8	2	5	6	8	3	5	10	6	10	7

$\Sigma x = 248$; $\Sigma y = 97$; $\Sigma x^2 = 4856$; $\Sigma y^2 = 731$; $\Sigma xy = 1825$

(d) During a week in which Dorothy makes 18 visits, how many people do you predict will buy insurance from her?

5. *Marketing: Coupons* Each box of Healthy Crunch breakfast cereal contains a coupon entitling you to a free package of garden seeds. At the Healthy Crunch home office, they use the weight of incoming mail to determine how many of their employees are to be assigned to collecting coupons and mailing out seed packages on a given day. (Healthy Crunch has a policy of answering all its mail on the day it is received.)

Let x = weight of incoming mail and y = number of employees required to process the mail in one working day. A random sample of 8 days gave the following data:

x (lb)	11	20	16	6	12	18	23	25
y (Number of employees)	6	10	9	5	8	14	13	16

$\Sigma x = 131$; $\Sigma y = 81$; $\Sigma x^2 = 2435$; $\Sigma y^2 = 927$; $\Sigma xy = 1487$

(d) If Healthy Crunch receives 15 lb of mail, how many employees should be assigned mail duty?

6. *Focus Problem: Changing Population and Crime Rate* Let x be a random variable representing percentage change in neighborhood population in the past few years, and let y be a random variable representing crime rate (crimes per 1000 population). A random sample of six Denver neighborhoods gave the following information (Source: *Neighborhood Facts,* The Piton Foundation).

x	29	2	11	17	7	6
y	173	35	132	127	69	53

$\Sigma x = 72$; $\Sigma y = 589$; $\Sigma x^2 = 1340$; $\Sigma y^2 = 72{,}277$; $\Sigma xy = 9499$

(d) For a neighborhood with $x = 12\%$ change in population in the past few years, predict the change in the crime rate (per 1000 residents).

DATA HIGHLIGHTS: GROUP PROJECTS

Break into small groups and discuss the following topics. Organize a brief outline in which you summarize the main points of your group discussion.

Scatter diagrams! Are they really useful? Scatter diagrams give a first impression of a data relationship and help us assess whether a linear relation provides a reasonable model for the data. In addition, we can spot *influential points*. A data point with an extreme x value can heavily influence the position of the least-squares line. In this project, we look at data sets with an influential point.

x	1	4	5	9	10	15
y	3	7	6	10	12	4

(a) Compute r and b, the slope of the least-squares line. Find the equation of the least-squares line, and sketch the line on the scatter diagram.

(b) Notice the point boxed in blue in Figure 4-11. Does it seem to lie away from the linear pattern determined by the other points? The coordinates of that point are (15, 4). Is it an influential point? Remove that point from the model and recompute r, b, and the equation of the least-squares line. Sketch this least-squares line on the diagram. How does the removal of the influential point affect the values of r and b and the position of the least-squares line?

(c) Consider the scatter diagram of Figure 4-12. Is there an influential point? If you remove the influential point, will the slope of the new least-squares line be larger or smaller than the slope of the line from the original data? Will the correlation coefficient be larger or smaller?

FIGURE 4-11

Scatter Diagram

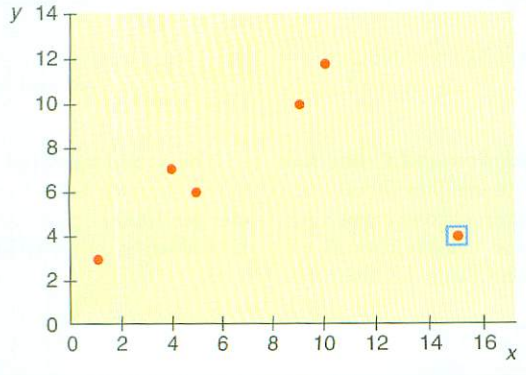
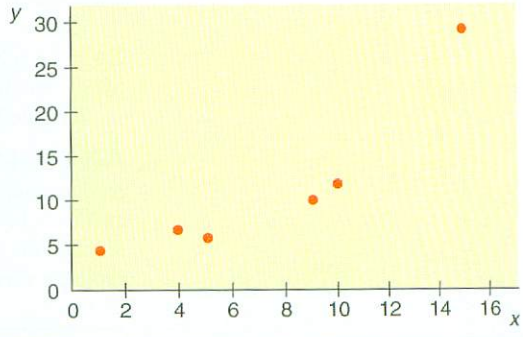


FIGURE 4-12

Scatter Diagram



LINKING CONCEPTS: WRITING PROJECTS

Discuss each of the following topics in class or review the topics on your own. Then write a brief but complete essay in which you summarize the main points. Please include formulas and graphs as appropriate.

1. What do we mean when we say that two variables have a strong positive (or negative) linear correlation? What would a scatter diagram for these variables look like? Is it possible that two variables could be strongly related somehow, but have a low *linear* correlation? Explain and draw a scatter diagram to demonstrate your point.

2. What do we mean by the least-squares criterion? Give a very general description of how the least-squares criterion is involved in the construction of the least-squares line. Why do we say the least-squares line is the "best-fitting" line for the data set?

3. Use the Internet or go to the library and find a magazine or journal article in your field of major interest to which the content of this chapter could be applied. List the variables used, method of data collection, and general type of information and conclusions drawn.

Using Technology

TI-84PLUS/TI-83PLUS • EXCEL • MINITAB • SPSS

Simple Linear Regression (one explanatory variable)

APPLICATION

The data in this section are taken from this reference:

King, Cuchlaine A. M. *Physical Geography.* Oxford: Basil Blackwell, 1980, 77–86, 196–206. Reprinted with permission of Basil Blackwell Limited, Oxford, England.

Throughout the world, natural ocean beaches are beautiful sights to see. If you have visited natural beaches, you may have noticed that when the gradient or dropoff is steep, the grains of sand tend to be larger. In fact, a manmade beach with the "wrong" size granules of sand tends to be washed away and eventually replaced when the proper size grain is selected by the action of the ocean and the gradient of the bottom. Since manmade beaches are expensive, grain size is an important consideration.

In the data that follow, x = median diameter (in millimeters) of granules of sand, and y = gradient of beach slope in degrees on natural ocean beaches.

x	y
0.17	0.63
0.19	0.70
0.22	0.82
0.235	0.88
0.235	1.15
0.30	1.50
0.35	4.40
0.42	7.30
0.85	11.30

1. Find the sample mean and standard deviation for x and y.
2. Make a scatter plot. Would you expect a moderately high correlation and a good fit for the least-squares line?
3. Find the equation of the least-squares line, and graph the line on the scatter plot.
4. Find the correlation coefficient r and the coefficient of determination r^2.
5. Suppose that you have a truckload of sifted sand in which the median size of granules is 0.38 mm. If you want to put this sand on a beach and you don't want the sand to wash away, then what does the least-squares line predict for the angle of the beach? *Note:* Heavy storms that produce abnormal waves may also wash out the sand. However, in the long run, the size of sand granules that remain on the beach or that are brought back to the beach by long-term wave action are determined to a large extent by the angle at which the beach drops off.
6. Suppose you now have a truckload of sifted sand in which the median size of the granules is 0.45 mm. Repeat Problem 5.

Technology Hints

TI-84Plus/TI-83Plus

Be sure to set **DiagnosticOn** (under **Catalog**).

(a) Scatter diagram: Use **STAT PLOT**, select the first type, and use **ZOOM** option **9:ZoomStat**.

(b) Least-squares line and r: Use **STAT, CALC**, option **8:LinReg(a + bx)**.

(c) Graph least-squares line and predict: Press **Y=**. Then, under **VARS**, select **5:Statistics**, then select **EQ**, and finally select item **1:RegEQ**. Press **Enter**. This sequence of steps will automatically set Y_1 = your regression equation. Press **GRAPH**. To find a predicted value when the graph is showing, press the **CALC** key and select item **1:Value**. Enter the x value and the corresponding y value will appear.

Excel

(a) Scatter plot, least-squares line, r^2: Use **Chart wizard**. Select **scatter plot**. Once the plot is displayed, *right* click on any data point. Select **trend line**. Under options, check display line and display r^2.

(b) Prediction: Use the paste function f_x ➤ **Statistical** ➤ **Forecast**.

(c) Coefficient r: Use f_x ➤ **Statistical** ➤ **Correl**.

Minitab

Scatter plot, least-squares line, r^2: Use the menu selection **Stat** ➤ **Regression** ➤ **Fitted line plot**.

SPSS

SPSS offers several options for finding the correlation coefficient r and the equation of the least-squares line. First enter the data in the data editor and label the variables appropriately in the variable view window. Use the menu choices **Analyze** ➤ **Regression** ➤ **Linear** and select dependent and independent variables. The output includes the correlation coefficient, the constant (y=intercept), and the coefficient of the dependent variable.